Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER

Minimalist models for protein folding and design

Teresa Head-Gordon* and Scott Brown

Protein folding research during the past decade has emphasized the dominant role of native state topology in determining the speed and mechanism of folding for small proteins; this has been illustrated by simulations using minimalist protein models. The advantages of minimalist protein models lie in their ability to rapidly collect meaningful statistics about folding pathways and kinetics, their ease of characterization with coarse-grained order parameters and their concentration on the essential physics of the problem to connect with experimental observables for a target protein. The maturation of experimental protein folding has driven the need for more quantitative protein simulations to better understand the balance between sequence details and fold topology. In the past year, we have seen the emergence of more complex minimalist models, ranging from all-atom Gō potentials to coarse-grained bead models in which Gō interactions are replaced or supplemented by more physically motivated potentials. The reduced computational cost at the coarse-grained level of abstraction will potentially enable both folding studies on a genomic scale and systematic application in protein design.

Addresses
Department of Bioengineering, University of California, Berkeley, CA 94720, USA
*e-mail: TLHead-Gordon@lbl.gov

Current Opinion in Structural Biology 2003, **13**:160–167

This review comes from a themed issue on
Theory and simulation
Edited by Charles L Brooks III and David A Case

0959-440X/03/$ – see front matter

DOI 10.1016/S0959-440X(03)00030-7

Abbreviations
MJ Miyazawa and Jernigan
SH Src homology

Introduction

Considerable physical insight into the protein folding process has been gained from advances in theoretical perspectives [1]. These are often explored by computer simulation with minimalist (suppression of significant computational complexity) protein models [1–3]. Energy landscape theories and supporting simulations have suggested that several thermodynamic measures, such as a smooth energy landscape [4], a large energy gap between native and misfolded free energy basins [4,5], and the σ-parameter (which measures the concomitant formation of native interactions with polymer collapse [6]), can be correlated with the fast folding kinetics of native protein sequences.

Early illustrations of the energy landscape theory involved simulations of the protein chain as a string of single-site beads with interactions that favor native state contacts of the target fold topology, also known as Gō potentials [1–3]. Minimalist Gō models minimize energetic roughness (traps) on the free energy surface and are topologically 'frustrated' only in folding to a particular three-dimensional shape. They provide a sufficient protein model for explaining why native sequences fold more rapidly and more reliably relative to poorly designed or arbitrary heteropolymer sequences [1,3–6]; these conclusions extend to the possibility that evolution has evolved sequences that favor fast folding [7]. They qualitatively reproduce differences in the folding kinetics of small and large proteins [8], and have been used to clarify the role of native state topology and minimal energetic frustration in the determination of the rate and mechanism of folding [9,10•].

Because Gō models avoid the more difficult aspect of the protein folding problem — its dependence on amino acid sequence — it is not surprising that these idealized models lack a quantitative connection to experiment in some cases [11••,12••]. For example, systematic experimental folding studies of approximately 20 small proteins have shown that, although they all fold by a two-state mechanism, their individual folding timescales vary over many decades [13] — something that is not reproduced by Gō models. An increase in model complexity is required to distinguish differences in the folding mechanisms of proteins with identical topologies, such as the classic example of proteins L and G. The presence or absence of early kinetic intermediates in folding is controversial in the experimental domain [14••,15••], and could be addressed by predictive protein folding simulations in which the model is fully characterizable.

Better connection to experimental folding can and will be made by increasing the accessible timescales of atomic-detail protein and solvent simulations [16•,17,18•]. An exciting development concerning the sampling problem for more complex potentials is being realized by new computing paradigms such as world-wide distributed computing [18•] and new hardware architectures from IBM's Blue Gene project [19•] that simulate the full protein folding event of populations of trajectories. These efforts have provided a more detailed connection to experimental studies [20•], will help validate empirical forcefields and drive the development of new

analysis techniques that cope with unprecedented volumes of data.

However, a model that contains this level of complexity is still too computationally expensive to be feasible given the amount of statistics required to fully characterize the thermodynamics and kinetics of the folding process. This is especially true when we consider the challenge of simulating a large number of protein folds, protein sizes and sequence mutations [21•]. The advantages of minimalist models lie in their ability to rapidly collect meaningful statistics about folding pathways, kinetics and thermodynamics, and their ease of characterization with coarse-grained order parameters.

Why would minimalist models be expected to be a reasonable approximation to the folding of real proteins? Real proteins exhibit a greater variety and subtlety of interactions among the different amino acids, cooperative formation of secondary structure through backbone hydrogen bonding and very specific sidechain packing of the native state core. The answer, in part, lies in the important observation by Plaxco *et al.* [22] that the magnitude of the folding rate for simple two-state folders is strongly correlated with average sequence separation between contacting residues in the native state. This emphasis on native topology is something that coarse-grained models do well by capturing the correct spatial distribution of local and nonlocal contacts, elements considered to be possibly the most important in governing the overall kinetics of protein folding [9,23•].

Currently, minimalist protein models are evolving toward making better connections to experiment by adding more chemical detail and physically motivated interactions. A number of studies have recently examined all-atom Gō potentials that have demonstrated an increase in folding cooperativity due to better sidechain packing [24••,25•]. These all-atom Gō potentials are one way to examine the more delicate balance between energetic and topological frustration for proteins in the ubiquitin and SH3 fold classes, for example. Alternatively, several research groups have considered sequence-dependent bead models, in which the Gō interactions are replaced by potentials of mean force that are derived from physical principles [26–29,30•–32•,33••,34,35], or Miyazawa and Jernigan (MJ) statistical potentials [36] laid over Gō interactions [37•]. Given the importance of aqueous solvent in protein folding and stability, minimalist protein models now place greater emphasis on residue interactions with more explicit features of hydration forces [38,39•].

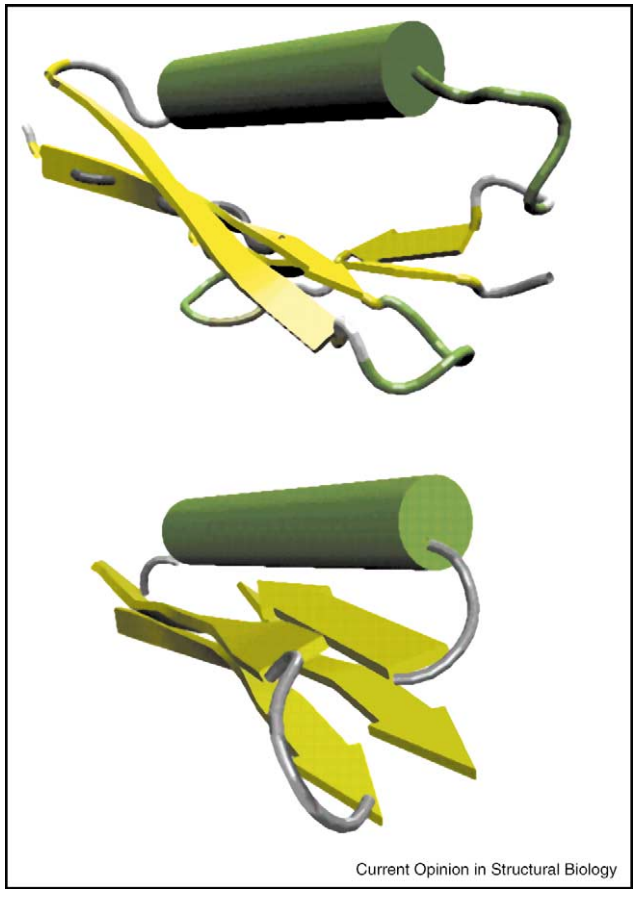
Here, we review progress in the development of minimalist models and their application to current research issues in protein folding. We conclude with an outlook of their promise in the wider context of the design and folding of many sequences and topologies.

The balance between topological and energetic frustration

Sequence-independent Gō models have minimal energetic frustration and therefore are inappropriate for explaining differences in mechanism or folding speed for proteins with low sequence identity but high structural homology. A classic example of this more delicate balance between energetic and topological frustration is found for proteins G and L, small proteins that adopt a ubiquitin fold consisting of a central α helix packed against a mixed β sheet composed of two β-hairpin structures (Figure 1). Experiments on proteins G and L have established that they fold by different pathways [40–44,45•]. Protein L folds through a polarized transition state, whereby the first β hairpin forms with the second β hairpin unstructured. By contrast, protein G folds through a transition state with purported rate-limiting formation of the second β hairpin.

A current note of discord in the experimental folding community is the existence of early intermediates in folding, that is, free energy barriers that precede the

Figure 1



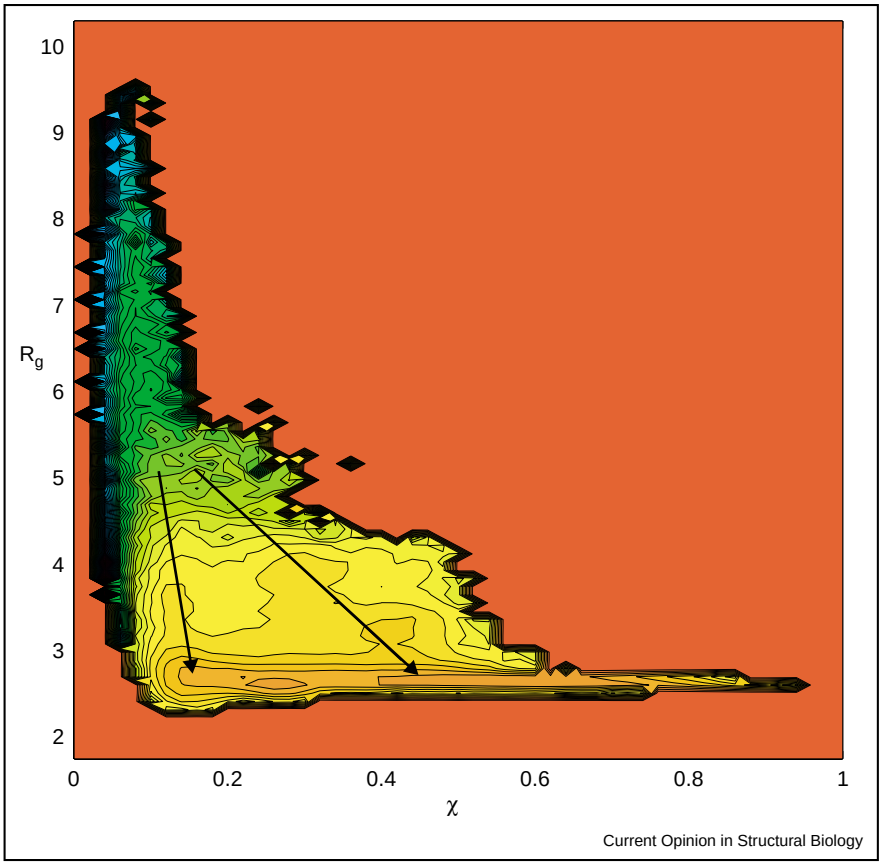
Minimalist model of the native state topology of protein L (bottom) and the NMR solution structure (top) [64], showing the similar arrangement of secondary and tertiary structure.

rate-limiting nucleation barrier of the folding reaction [15••]. Although there is agreement that the folding of protein L appears to be purely two state [43,44], protein G shows evidence of an early intermediate along the folding pathway [41], although this result has been contested as a problem of the suspect interpretation of ultrafast folding events in general [15••].

Recent simulations of coarse-grained models of proteins L and/or G (Figure 1) have been reported by several research groups using different enhanced forms of minimalist models. Shimada and Shakhnovich [24••] have used ensemble dynamics to characterize the kinetics of protein G using an all-atom Gō potential. Karanicolas and Brooks [37•] used a Gō potential bead model supplemented by sequence-dependent MJ statistical potentials to differentiate the folding of proteins G and L. They found the origin of asymmetry in the folding of proteins L and G to be in concurrence with Nauli *et al.* [45•], who used a computer-based design strategy to re-engineer the protein G sequence to include more stabilizing interactions for the first β-hairpin turn, producing a protein more faithful to the folding of protein L. Brown and Head-Gordon [33••] have simulated the folding of proteins G and L using an off-lattice bead model developed as the first sequence-dependent minimalist model for α/β topologies, in which the tertiary Gō interactions are replaced by potentials of mean force derived from physical principles [30•–32•,33••].

The consensus of the two extended minimalist models used to examine the kinetics of protein G [24••,33••] determined that protein G folds through multiple pathways and, in addition, some or all of these pathways involve an intermediate. A portion of folding trajectories follow a fast pathway consistent with collapse concomitant with native structure formation (which is equivalent mechanistically, but not in full detail, to that found for protein L) (Figure 2). Others followed a slow pathway (best fit to a double exponential) involving an early event of strong compaction with some structuring of the second β hairpin, and a long timescale for correcting the consequences of

Figure 2



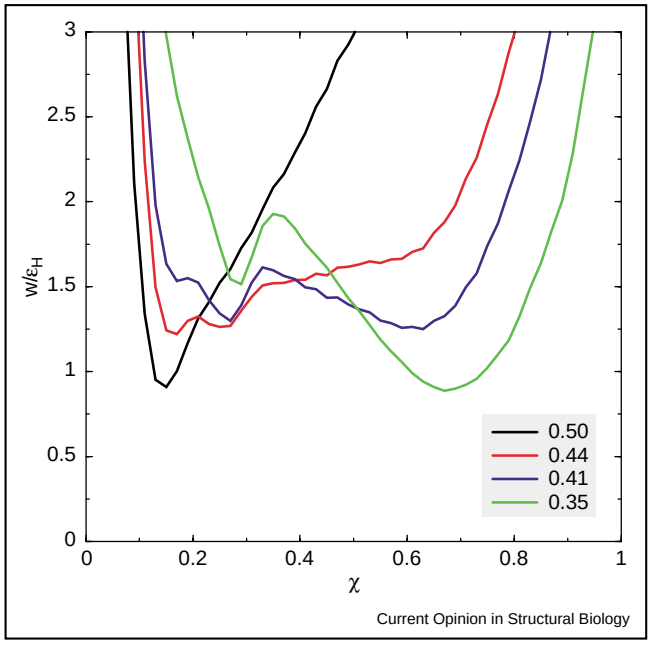
Free energy at the folding temperature as a function of radius of gyration (R_g) and native state similarity (χ) for protein G. Two folding pathways are present. The fast pathway corresponds to a collapse-concomitant folding pathway (arrow on right), whereas the slow pathway (arrow on left) corresponds to rapid non-native collapse with a structured second β hairpin and a longer process of finding the native structure. The contour lines are spaced at intervals of k_bT, with blue to red representing high to low free energy values.

non-native collapse (Figure 2). In the study reported in [33••], protein G folds slower than protein L by a factor of two, qualitatively consistent with experiment.

The theoretical studies strongly emphasize that the choice of reaction coordinate for monitoring folding progress is important for the observation of intermediates [24••,33••,46]. Shimada and Shakhnovich [24••] showed that, when folding is monitored by using burial of the lone tryptophan in protein G as the reaction coordinate, the ensemble kinetics is single exponential (as was observed in the all-atom simulations reported in [46]), whereas alternative reaction coordinates revealed the presence of intermediates along the multiple pathways.

The problem of determining a good reaction coordinate is illustrated in Figure 3, which shows the potential of mean force for protein G as a function of native state similarity in going from the unfolded to folded states as a function of temperature (the folding temperature for the protein G sequence is T = 0.41 [reduced units]) [33••]. These results might be interpreted as a shift in unfolded population with increasing native state stability, in agreement with Qi *et al.* [47], and might provide an alternative interpretation of ultrafast folding experiments that would be consistent with two-state folding.

Figure 3



Potential of mean force versus native state similarity as a function of temperature for the folding of protein G for T = 0.50 (black), T = 0.44 (red), T = 0.41 (blue) and T = 0.35 (green). The folding temperature is T = 0.41 (reduced units) and the contour lines are spaced at intervals of k_bT. Based on this projection, we might conclude that there is a shift in the unfolded population as we approach folding conditions. There is also evidence of a small barrier.

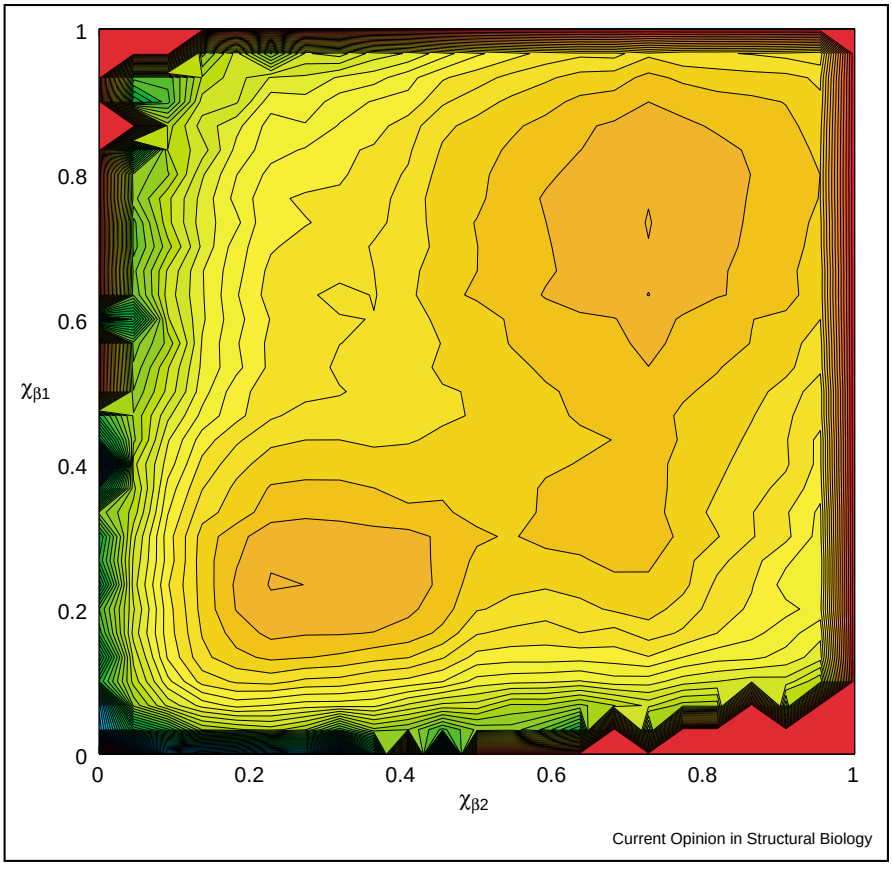
Different projections onto the free energy surface that follow β-hairpin reaction coordinates show initial formation of the second β hairpin, followed by formation of the first β hairpin (Figure 4). In all projections (Figures 2–4), we see the presence of a small barrier near folding conditions, consistent with the fact that the entire population of kinetic runs was best fit to a double exponential and consistent with formation of an intermediate (Figure 5) [33••]. The small free energy barriers observed in these cuts through the free energy surface suggest to us that the full reaction coordinate is not fully revealed and that a more complicated reaction coordinate may be required [33••]. This again emphasizes that the choice of reaction coordinate used experimentally is very important, if different conclusions concerning the presence of intermediates are to be avoided, as was found to be the case for re-examination of the presence of an intermediate in ubiquitin [14••].

The role of solvation in protein folding

It is widely appreciated that water plays an important role in governing the forces that control protein structure and stability [48]. The strong hydration forces that are responsible for hydrophobic attraction and stabilization of a protein's native core are expected to also play an important role in governing how the protein folds quickly to the proper folded state. Most minimalist models have concentrated on interactions that do not include explicit residue–water and water–water interactions. There is of course implicit incorporation of solvation in physical potentials through sticky hydrophobic interactions. However, the use of pairwise contact potentials neglects two prominent features of hydration forces: their many-body nature and their potentially longer-range effects. Recently, several research groups have attempted to extend minimalist models to include more explicit investigation of the profound effect of hydration forces on protein folding [39•,49•,50].

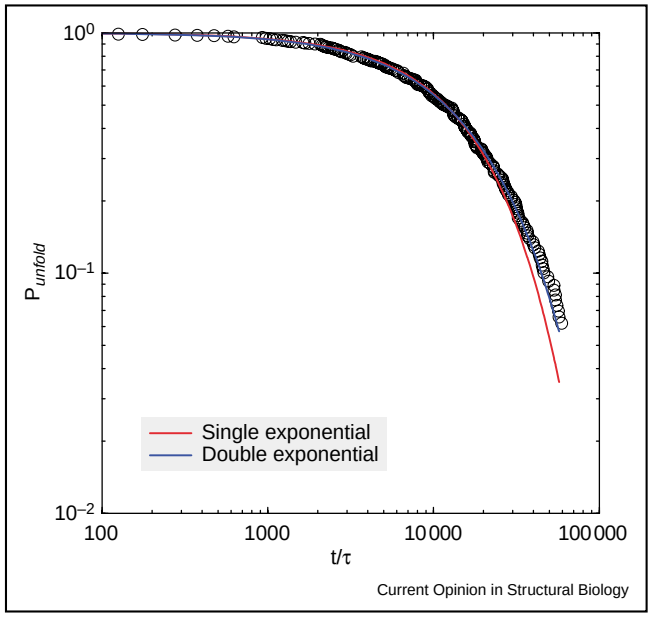
Cheung *et al.* [39•] have recently extended Gō potential bead models to include an explicit solvation interaction that favors direct contacts between beads, but also has a barrier separating this minimum from a weak secondary 'water-separated' minimum. This solvation potential of mean force is based on previous work from the chemical-physics community that quantifies hydration for small hydrophobic entities in water [51,52] and that has been proposed to be the correct solvation physics relevant in protein folding [50,53,54]. The observation of water impregnation late in folding and the requirement of overcoming a desolvation barrier to reach the native state have been observed in a large number of simulations, including the refolding of a β-hairpin fragment of protein G [55], and all-atom simulations of protein G [56] and the src SH3 domain [57•]. The paper by Cheung *et al.* [39•] nicely demonstrated that the desolvation barrier plays a major role in the mechanism of SH3 domain folding, in which a near-native intermediate with a partially solvated

Figure 4



Free energy at the folding temperature as a function of χ_{β_1} and χ_{β_2} for protein G. The contour lines are spaced at intervals of k_bT, with blue to red representing high to low free energy values. The low free energy path corresponds to the formation of native β-hairpin 2 with non-native β-hairpin 1.

Figure 5



Fraction of unfolded states versus time at the folding temperature T = 0.41 for protein G. Circles correspond to simulated data, the blue line is a bi-exponential fit to the data and the red line is the fit to single-exponential kinetics. The figure shows that the kinetics of protein G are best fit to a double exponential.

hydrophobic core is reached before final expulsion of water molecules to reach the native state. This partially denatured state appears to be consistent with residues observed experimentally to be partially hydrated in the vicinity of the core [39•,58].

Recent simulations have made it increasingly apparent that hydration forces can be strongly non-pairwise additive [50,53]. A lattice protein folding study investigated the effect of adding a multibody description of hydration to a simple two-flavor protein model [50]. Sequences in the hydrated model were more frequently found to have unique ground states, to fold faster and to fold with more cooperativity than sequences in the corresponding model without solvation terms. These results indicate that the multibodied nature of hydration is a counterpart to amino

acid diversity and packing, which in turn gives rise to a more cooperative folding transition [50].

Conclusions

The advantages of minimalist protein models lie in their ability to rapidly collect meaningful statistics about folding pathways and kinetics, their ease of characterization with coarse-grained order parameters and their concentration on the essential physics of the problem to connect with experimental observables for a given target protein. Although Gō bead models have emphasized that native interaction biases in the free energy landscape guide the unfolded chain to the native state, these models avoid the original and more difficult problem of what are the actual physical interactions that give rise to these biases in the real free energy surface. The deficiencies in these models have been recognized and, over the past year, have been supplemented by enough additional complexity to help understand the delicate balance between energetic and topological frustration, and the influence of aqueous hydration on folding. This is important for several reasons when we consider the future of protein folding simulation using minimalist models.

First, we need to consider the impact of genome sequencing projects, which provide an opportunity to understand biological systems at a whole new level of complexity [21•]. The first sweep through the genomic data has emphasized large sequence comparison studies within and between genomes to infer the structure and function of new proteins that bear analogy to existing proteins whose structure or biochemistry is known. In some cases, however, this inference approach will not be useful if no analogs exist or if the chemistry of the new sequence is different enough from that of the analog to exhibit changes in folding, structure, dynamics and/or function. Biophysical approaches should be complementary to bioinformatics for completing the annotation and minimalist models with reduced complexity should contribute here. For example, a 'designability principle' has emerged based on minimalist hydrophobic-polar (HP) and MJ protein potentials that might help explain the organization of genomic sequences into fold families or superfolds [59,60,61•,62].

Understanding protein self-assembly outside the context of structural biology found in nature requires a better understanding of the underlying physical interactions for successful design of new materials. It has been recently reported that the diversity of designed sequences is primarily determined by a structure's overall fold [63•], in accordance with simple protein folding models [59,60,61•,62]. Again, the strong role of topology suggests that further practical success in protein design might benefit from fully characterizing the thermodynamic and kinetic measures of foldability using the enhanced minimalist models discussed here.

Finally, biophysical theories and models of folding have not yet made a connection to protein function and its variation across genomes. Could we correlate a slow folding but very stable protein with a ubiquitous house-keeping function or proteins that fold with intermediates as potential candidates for chaperonin assistance or indicators of disease? We will end with the provocative suggestion that minimalist models, because they are completely characterizable and tractable, could be extended from their original goal of understanding the general rules of folding to learning the general rules of function on a genomic scale.

Acknowledgements

TH-G would like to acknowledge financial support for our minimalist model work from UC Berkeley.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Onuchic JN, Luthey-Schulten Z, Wolynes PG: **Theory of protein folding: the energy landscape perspective**. *Annu Rev Phys Chem* 1997, **48**:545-600.

2. Go N: **Protein folding as a stochastic process.** *J Stat Phys* 1983:413-423.

3. Dill KA, Chan HS: **From Levinthal to pathways to funnels**. *Nat Struct Biol* 1997, **4**:10-19.

4. Bryngelson JD, Wolynes PG: **Intermediates and barrier crossing in a random energy model (with applications to protein folding)**. *J Phys Chem* 1989, **93**:6902-6915.

5. Sali A, Shakhnovich E, Karplus M: **How does a protein fold**. *Nature* 1994, **369**:248-251.

6. Klimov DK, Thirumalai D: **Factors governing the foldability of proteins**. *Proteins* 1996, **26**:411-441.

7. Mirny LA, Abkevich VI, Shakhnovich EI: **How evolution makes proteins fold quickly**. *Proc Natl Acad Sci USA* 1998, **95**:4976-4981.

8. Cieplak M, Hoang TX, Li MS: **Scaling of folding properties in simple models of proteins**. *Phys Rev Lett* 1999, **83**:1684-1687.

9. Shea JE, Onuchic JN, Brooks CL: **Exploring the origins of topological frustration: design of a minimally frustrated model of fragment B of protein A**. *Proc Natl Acad Sci USA* 1999, **96**:12512-12517.

10. • Clementi C, Jennings PA, Onuchic JN: **Prediction of folding mechanism for circular-permuted proteins**. *J Mol Biol* 2001, **311**:879-890.

This paper substantiates the role of topological frustration in a joint experimental/theoretical study of circularly permuted proteins.

11. •• Koga N, Takada S: **Roles of native topology and chain-length scaling in protein folding: a simulation study with a Go-like model**. *J Mol Biol* 2001, **313**:171-180.

This paper analyzes folding rate constants and characteristics of the transition state ensemble and the denatured states in terms of native topology and chain length for 18 different proteins using a sequence-independent Gō model. The transition state ensemble of the model is consistent with experimental Φ-values for approximately half of the proteins, indicating that sequence dependence is important to folding for the other half.

12. •• Plaxco KW, Simons KT, Ruczinski I, David B: **Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics**. *Biochemistry* 2000, **39**:11177-11183.

The authors make the important observation that the magnitude of the folding rate for simple two-state folders is strongly correlated with the average sequence separation between contacting residues in the native state. This emphasizes the role of native topology in determining the kinetics of folding, at least for small proteins.

13. Makarov DE, Keller CA, Plaxco KW, Metiu H: **How the folding rate constant of simple, single-domain proteins depends on the number of native contacts**. *Proc Natl Acad Sci USA* 2002, **99**:3535-3539.

14. •• Qin Z, Ervin J, Larios E, Gruebele M, Kihara H: **Formation of a compact structured ensemble without fluorescence signature early during ubiquitin folding**. *J Phys Chem* 2003, in press.
This comprehensive analysis of the folding of ubiquitin supports the presence of an early intermediate, while at the same time emphasizing caution in experimental analysis of kinetics using fluorescence signatures. It also strongly implies that care must be taken in the choice of experimental reaction coordinate followed to observe protein intermediates.

15. •• Krantz BA, Mayne L, Rumbley J, Englander SW, Sosnick TR: **Fast and slow intermediate accumulation and the initial barrier mechanism in protein folding**. *J Mol Biol* 2002, **324**:359-371.
The authors argue that the finding of early intermediates (extra barrier(s) that precede the rate-limiting barrier to folding) is an artifact of interpreting ultrafast experimental techniques.

16. • Brooks CL: **Protein and peptide folding explored with molecular simulations**. *Acc Chem Res* 2002, **35**:447-454.
Molecular simulations with emphasis on models with full atomic detail of polypeptide and solvent.

17. Duan Y, Kollman PA: **Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution**. *Science* 1998, **282**:740-744.

18. • Shirts M, Pande VS: **Computing - Screen savers of the world unite!** *Science* 2000, **290**:1903-1904.
The authors describe using world-wide distributed computing in the spirit of the SETI project to overcome the sampling problem for high-resolution protein folding models.

19. • Allen F, Almasi G, Andreoni W, Beece D, Berne BJ, Bright A, Brunheroto J, Cascaval C, Castanos J, Coteus P *et al.*: **Blue Gene: a vision for protein science using a petaflop supercomputer**. *IBM Systems Journal* 2001, **40**:310-327.
This paper describes the design of special-purpose computer hardware and software for application to protein science.

20. • Snow CD, Nguyen N, Pande VS, Gruebele M: **Absolute comparison of simulated and experimental protein-folding dynamics**. *Nature* 2002, **420**:102-106.
A comparison of kinetics determined from experiment and simulation. The folding@home simulation predicted that protein BBA5 would fold in 6 μs, whereas experiment revealed an actual folding time of 7.5 μs.

21. • Head-Gordon T, Wooley JC: **Computational challenges in structural and functional genomics**. *IBM Systems Journal* 2001, **40**:265-296.
A blueprint for computational biology research over the next several decades.

22. Plaxco KW, Simons KT, Baker D: **Contact order, transition state placement and the refolding rates of single domain proteins**. *J Mol Biol* 1998, **277**:985-994.

23. • Nymeyer H, Socci ND, Onuchic JN: **Landscape approaches for determining the ensemble of folding transition states: success and failure hinge on the degree of frustration**. *Proc Natl Acad Sci USA* 2000, **97**:634-639.
This paper shows that the connection between the folding kinetics and the thermodynamics of the transition state ensemble depends on the degree of energetic frustration. It also shows that interpretation of experimentally measured Φ-values as changes in free energy differences for a simple transition state ensemble is accurate and useful only for those proteins that are minimally frustrated and whose folding can be characterized by simple reaction coordinates.

24. •• Shimada J, Shakhnovich EI: **The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation**. *Proc Natl Acad Sci USA* 2002, **99**:11175-11180.
A recent extension of minimalist bead models to all-atom Gō potentials to characterize the folding of protein G. This paper shows that the presence of intermediates in folding may be lost when doing ensemble dynamics.

25. • Clementi C, Garcia A, Onuchic JN: **Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: an all-atom representation study**. *J Mol Biol* 2003, **326**:933-954.
A study using an all-atom Gō potential to characterize the folding of proteins G and L. The addition of sidechain packing results in greater cooperativity of folding relative to sequence-independent Gō models.

26. Honeycutt JD, Thirumalai D: **Metastability of the folded states of globular proteins**. *Proc Natl Acad Sci USA* 1990, **87**:3526-3529.

27. Guo ZY, Thirumalai D, Honeycutt JD: **Folding kinetics of proteins - a model study**. *J Chem Phys* 1992, **97**:525-535.

28. Guo Z, Thirumalai D: **Kinetics and thermodynamics of folding of a *de novo* designed four-helix bundle protein**. *J Mol Biol* 1996, **263**:323-343.

29. Sorenson JM, Head-Gordon T: **Redesigning the hydrophobic core of a model beta-sheet protein: destabilizing traps through a threading approach**. *Proteins* 1999, **37**:582-591.

30. • Sorenson JM, Head-Gordon T: **Matching simulation and experiment: a new simplified model for simulating protein folding**. *J Comput Biol* 2000, **7**:469-481.
The first physico-chemical model for the α/β topology, based on the ubiquitin fold family.

31. • Sorenson JM, Head-Gordon T: **Protein engineering study of protein L by simulation**. *J Comput Biol* 2002, **9**:35-54.
The authors validated the folding of the minimalist protein L/G model by comparison with sequence mutation data and Φ-value analysis. This paper emphasizes the ability of minimalist models to fully analyze and characterize folding.

32. • Sorensen JM, Head-Gordon T: **Toward minimalist models of larger proteins: a ubiquitin-like protein**. *Proteins* 2002, **46**:368-379.
The extension of the physico-chemical model for protein L/G to investigate the folding of ubiquitin.

33. •• Brown S, Head-Gordon T: **Sequence design for determining proteins that fold by alternative folding mechanisms**. *Protein Sci* 2003, in press.
A physico-chemical bead model and sequence design strategy can be made to distinguish between the kinetics and mechanisms of protein L and protein G. This paper also shows that the reaction coordinate for protein G folding must be carefully chosen to see stable early intermediates.

34. Hao MH, Scheraga HA: **Designing potential energy functions for protein folding**. *Curr Opin Struct Biol* 1999, **9**:184-188.

35. Takada S, Luthey-Schulten Z, Wolynes PG: **Folding dynamics with nonadditive forces: a simulation study of a designed helical protein and a random heteropolymer**. *J Chem Phys* 1999, **110**:11616-11629.

36. Miyazawa S, Jernigan RL: **Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading**. *J Mol Biol* 1996, **256**:623-644.

37. • Karanicolas J, Brooks CL: **The origins of asymmetry in the folding transition states of protein L and protein G**. *Protein Sci* 2002, **11**:2351-2361.
The development of a minimalist protein model that includes physical interactions using an MJ statistical potential on top of a Gō potential to realize a sequence-dependent model for characterizing proteins G and L.

38. Sorenson JM, Hura G, Soper AK, Pertsemlidis A, Head-Gordon T: **Determining the role of hydration forces in protein folding**. *J Phys Chem B* 1999, **103**:5413-5426.

39. • Cheung MS, Garcia AE, Onuchic JN: **Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse**. *Proc Natl Acad Sci USA* 2002, **99**:685-690.
This paper describes the mechanism of folding of a simple SH3 model with explicit inclusion of a water reaction coordinate.

40. Gu HD, Yi QA, Bray ST, Riddle DS, Shiau AK, Baker D: **A phage display system for studying the sequence determinants of protein folding**. *Protein Sci* 1995, **4**:1108-1117.

41. Park SH, Oneil KT, Roder H: **An early intermediate in the folding reaction of the B1 domain of protein G contains a native-like core**. *Biochemistry* 1997, **36**:14277-14283.

42. Gu HD, Kim D, Baker D: **Contrasting roles for symmetrically disposed beta-turns in the folding of a small protein**. *J Mol Biol* 1997, **274**:588-596.

43. Plaxco KW, Millett IS, Segel DJ, Doniach S, Baker D: **Chain collapse can occur concomitantly with the rate-limiting step in protein folding**. *Nat Struct Biol* 1999, **6**:554-556.

44. Scalley ML, Yi Q, Gu HD, McCormack A, Yates JR, Baker D: **Kinetics of folding of the IgG binding domain of peptostreptoccocal protein L**. *Biochemistry* 1997, **36**:3373-3382.

45. • Nauli S, Kuhlman B, Baker D: **Computer-based redesign of a protein folding pathway**. *Nat Struct Biol* 2001, **8**:602-605.
Computational redesign and experimental verification of a switch in the folding pathway of protein G to favor the formation of the first β hairpin, similar to protein L.

46. Sheinerman FB, Brooks CL: **Calculations on folding of segment B1 of streptococcal protein G**. *J Mol Biol* 1998, **278**:439-456.

47. Qi PX, Sosnick TR, Englander SW: **The burst phase in ribonuclease A folding and solvent dependence of the unfolded state**. *Nat Struct Biol* 1998, **5**:882-884.

48. Kauzmann W: **Some factors in the interpretation of protein denaturation**. *Adv Protein Chem* 1959, **14**:1-59.

49. • Kaya H, Chan HS: **Solvation effects and driving forces for protein thermodynamic and kinetic cooperativity: how adequate is native-centric topological modeling?** *J Mol Biol* 2003, **326**:911-931.
This paper critically examines whether the use of Gō potentials, even when supplemented by physical hydration forces, is adequate for robust and predictive protein folding studies.

50. Sorenson JM, Head-Gordon T: **The importance of hydration for the kinetics and thermodynamics of protein folding: simplified lattice models**. *Fold Des* 1998, **3**:523-534.

51. Pratt LR, Chandler D: **Theory of the hydrophobic effect**. *J Chem Phys* 1977, **67**:3683-3704.

52. Hummer G, Garde S, Garcia AE, Paulaitis ME, Pratt LR: **Hydrophobic effects on a molecular scale**. *J Phys Chem B* 1998, **102**:10469-10482.

53. Rank JA, Baker D: **A desolvation barrier to hydrophobic cluster formation may contribute to the rate-limiting step in protein folding**. *Protein Sci* 1997, **6**:347-354.

54. Hura G, Sorenson JM, Glaeser RM, Head-Gordon T: **Solution X-ray scattering as a probe of hydration-dependent structuring of aqueous solutions**. *Perspectives in Drug Discovery and Design* 1999, **17**:97-118.

55. Pande VS, Rokhsar DS: **Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein G**. *Proc Natl Acad Sci USA* 1999, **96**:9062-9067.

56. Sheinerman FB, Brooks CL: **Molecular picture of folding of a small alpha/beta protein**. *Proc Natl Acad Sci USA* 1998, **95**:1562-1567.

57. • Shea JE, Onuchic JN, Brooks CL: **Probing the folding free energy landscape of the src-SH3 protein domain**. *Proc Natl Acad Sci USA* 2002, **99**:16064-16068.
An all-atom simulation study of the mechanism of SH3 folding using importance sampling, showing that the latest stages of folding involve the formation of the hydrophobic core through the expulsion of water molecules, consistent with the results described in [39•].

58. Zhang OW, Formankay JD: **NMR studies of unfolded states of an SH3 domain in aqueous solution and denaturing conditions**. *Biochemistry* 1997, **36**:3959-3970.

59. Li H, Helling R, Tang C, Wingreen N: **Emergence of preferred structures in a simple model of protein folding**. *Science* 1996, **273**:666-669.

60. Li H, Tang C, Wingreen NS: **Are protein folds atypical?** *Proc Natl Acad Sci USA* 1998, **95**:4987-4990.

61. • Li H, Tang C, Wingreen NS: **Designability of protein structures: a lattice-model study using the Miyazawa-Jernigan matrix**. *Proteins* 2002, **49**:403-412.
This theoretical study suggests that the most common folds of real proteins are the most atypical in the space of all possible structures and are more designable (defined as the number of sequences that have that structure as their unique lowest energy state), with enhanced thermodynamic stability. This paper confirms that these conclusions are sound whether using a hydrophobic-polar (HP) or MJ model.

62. Helling R, Li H, Melin R, Miller J, Wingreen N, Zeng C, Tang C: **The designability of protein structures**. *J Mol Graph Model* 2001, **19**:157-167.

63. • Larson SM, England JL, Desjarlais JR, Pande VS: **Thoroughly sampling sequence space: large-scale protein design of structural ensembles**. *Protein Sci* 2002, **11**:2804-2813.
The authors, using distributed computing, broadly explore sequence space with backbone flexibility to achieve large-scale protein design of structural ensembles.

64. Wikstrom M, Drakenberg T, Forsen S, Sjobring U, Bjorck L: **Three-dimensional solution structure of an immunoglobulin light chain-binding domain of protein l - comparison with the IgG-binding domains of protein G**. *Biochemistry* 1994, **33**:14011-14017.